# Fighting Poverty: Statistical Analysis for Identifying At-Risk Program Participants

Claire Danaher, Russ Davis, and Renee Sweeney

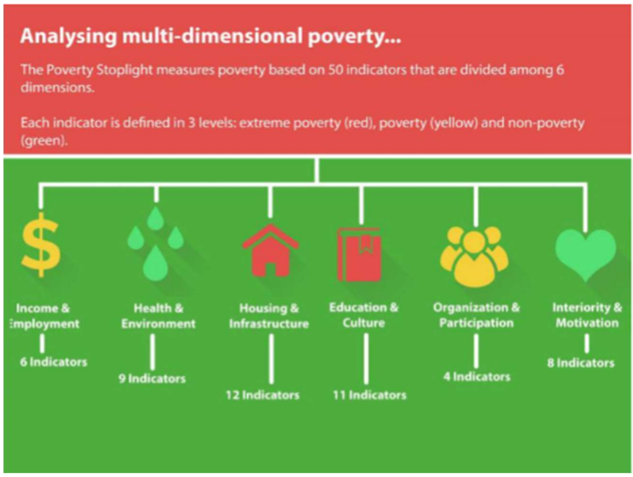
# 1 Abstract

Non-government organizations who work to reduce poverty in third world countries typically face the problem of efficient resource allocation. Since all families are different, they cannot always be evaluated equally in terms of need, and prioritizing certain causes of poverty becomes difficult. Given past data, classification models allow us to predict and infer which families are likely to struggle the most in overcoming poverty. In this paper, we present our methods and results for discriminant analysis, logistic regression, support vector machines, decision trees, and k-nearest neighbors on the Stoplight Program dataset from Fundación Paraguaya.

# 2 Introduction

Fundación Paraguaya is a non-governmental organization located in Paraguay. The goal of this organization is to support entrepreneurship as a path for eradicating poverty in Paraguay. One of the efforts organized by this foundation is the Poverty Stoplight Program (Stoplight Program). As described by the organization, the Stoplight program: “makes poverty ‘visible’ by dividing the [multidimensional poverty] model into 6 dimensions and 50 indicators, so that a poor person can visualize the ways in which poverty affects their own family.”

The figure below is featured in a brochure distributed by Fundación Paraguaya. The figure outlines the 6 poverty dimensions and 50 indicators.



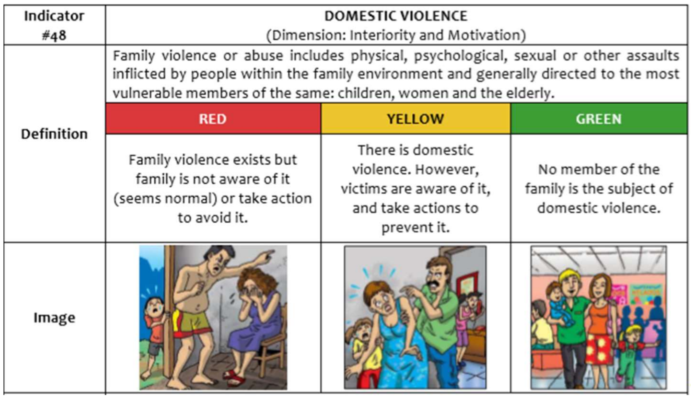
**Analysing multi-dimensional poverty...**

The Poverty Stoplight measures poverty based on 50 indicators that are divided among 6 dimensions.

Each indicator is defined in 3 levels: extreme poverty (red), poverty (yellow) and non-poverty (green).

[1]

For each participant in the program, program staff collects data on the 50 indicators. The indicators are displayed visually using both written verbal descriptions and images. The format of the survey asks participants to categorize their family as either: Red (for Extremely Poor), Yellow (for Poor), Green (for Not Poor) on each of the indicators. An example of one of the visualizations is provided below.



| Indicator #48 | DOMESTIC VIOLENCE<br>(Dimension: Interiority and Motivation) | | |
|---|---|---|---|
| Definition | Family violence or abuse includes physical, psychological, sexual or other assaults inflicted by people within the family environment and generally directed to the most vulnerable members of the same: children, women and the elderly. | | |
| | RED | YELLOW | GREEN |
| | Family violence exists but family is not aware of it (seems normal) or take action to avoid it. | There is domestic violence. However, victims are aware of it, and take actions to prevent it. | No member of the family is the subject of domestic violence. |
| Image | | | |

[2]

Our team was interested in analyzing the relationship between the poverty indicators of participants in the program and the likelihood that participation in the program would result in an improvement to a participant's overall poverty. More specifically, our analysis had the following goals:

[1] Growing with "El Mejor", Impact Measurement Report, Fundación Paraguaya, August 2015

- Develop a model that would predict whether a participant's state of poverty was unlikely to improve during participation in the program
- Conduct inferential analysis to try to better understand commonalities among participants whose poverty was not improved

The primary focus of our investigation is to allow the Stoplight program to optimize resource allocation by focusing on participants who are most likely to require additional help. Accordingly, while overall accuracy is important in model selection, we are also require the maximization of the true negative rate while minimizing the false negative rate.

# 3 Methods

## 3.1 Overview

The data collected by this program is primarily comprised of ordinal indicator variables with categories including extremely poor, poor, and not poor. The data was extremely high dimensional with 50 indicators and a few additional variables. Greater detail regarding the data cleaning process and modeling approaches are described below.

## 3.2 Data Cleaning

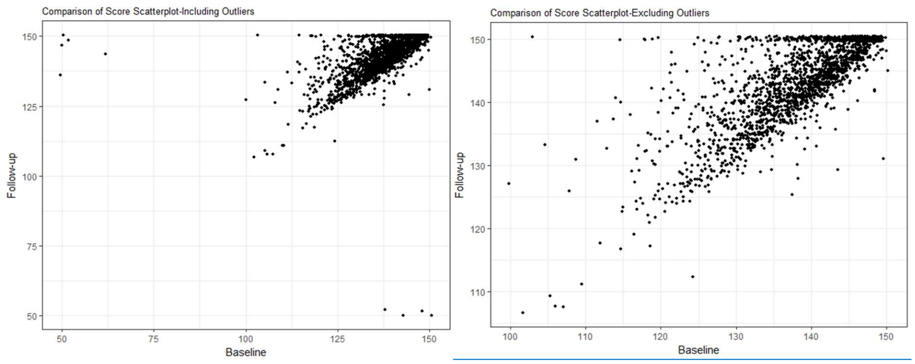
The raw data set provided by Fundación Paraguay included records for participants when they first entered the program(baseline data) and after participation in the program(follow-up data). Data points collected include the following for each participant:

- Unique Participant ID
- Advisor ID
- Branch Office ID
- Income
- Date
- Type of Client
- Area(Metropolitan/Rural)
- Baseline
- Fields for each of the 50 indicators
- Fields tracking the number of interactions with program personnel to improve each of the 50 indicators

As previously mentioned, the data included at least a baseline record for each participant and follow-up record(s) for many participants. Because our analysis was focused on improvement or lack thereof during participation in the program, we only considered individuals with follow-up records, and limited our analysis to the indicator values. The following data manipulation was required to complete this task:

- Baseline: The raw data set contained a baseline field which flagged the baseline record. However, this field was not consistently filled out. Therefore, the baseline was identified as the record earliest date for a given participant.
- Follow-Up: The data set contained multiple follow-up records for some participants. In these cases, the most recent record categorized as the follow-up record.
- Overall poverty score: Each record received an overall poverty score, which was the sum of all 50 indicators
- Improvement class: This variable was added to denote whether or not a participant's overall poverty score improved.
- Indicator values:  In some cases, indicator values of 0 were present, suggesting errors in data input.  In other cases, the indicator values from the baseline to the follow-up contained tremendous variation that was, again, indicative of errors in data input.  In both cases, these participants were not considered in our model to remove the influence of outliers.  Figure X shows the distribution of data before and after the removal of outliers.
- Aggregate dimension score: Some of the approaches detailed below require some dimension reduction and numerical variables rather than ordinal ones.  To achieve this, we aggregated scores of the indicators in each poverty dimension to obtain six dimensions for each observation each with a numerical score.

**Commented [1]:** Figure X



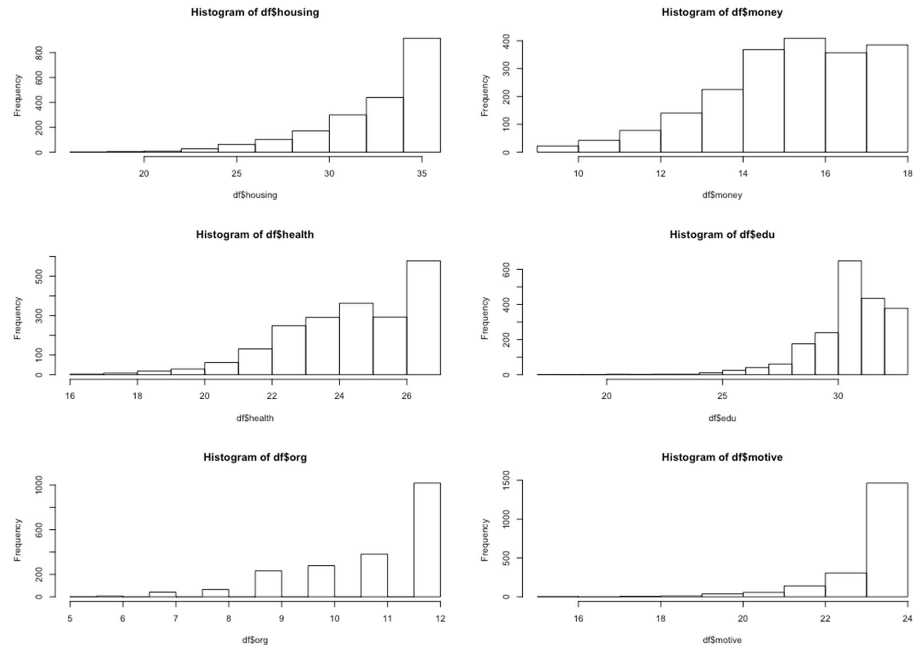
## 3.3 Data Challenges

There were four major challenges associated with this data set.

| Issue | Description |
|---|---|

| | |
|---|---|
| High Dimensional | 50 poverty indicators were considered. Records were provided for 2026 participants for whom complete records of both a baseline and a follow-up visit were provided. The ratio between instances and indicators was a little above 40 to 1. |
| Ordinal Data | The data collected asked participants to categorize themselves on a scale of poor to not poor. Ordinal data is challenging because there is an implied direction and distance but the distance between levels may not be consistent across indicators. |
| Small sample size for individuals with no improvement | The data provided has a much larger number of participants for which improvement was achieved as opposed to not achieved. |
| Skewed data | As shown in Figure X, the poverty dimensions were all skewed left. This represents a potential problem for any methodology that assumes an even distribution. |

**Commented [2]:** Figure X



## 3.4 Modeling

A number of different modeling approaches were explored for this analysis. The models selected for prediction purposes were those that would produce a binary predicted value(improve/didn't improve) and where a large number of ordinal predictor variables could be used.
Our team analyzed the data using the following approaches:

- Multiple Factor Analysis
- Classical Classification Approaches- QDA, LDA, and Logistic Regression
- Support Vector Machines- Linear, Polynomial, and Radial Kernels
- Trees- Boosting, Bagging and Random Forests

The results of these approaches are outlined in the following section.

# 4 Results

## 4.1 Dimension Reduction

The data provided was very high dimensional. The data is ordinal, meaning that Principal Component Analysis(PCA) was not a viable option. An alternative to PCA for categorical variables is Multiple Component Analysis(MCA). Rather than using covariance as in the case of PCA, MCA uses the entropy of the data set as the differentiation variable[2]. An exploration using multiple factor analysis was conducted to potentially identify differentiation factors among the data. For the purposes of this analysis, the indicators were grouped into along the 6 dimensions and then assigned a categorical variable depending upon the aggregate score. Participants in were grouped into three groups: bottom quartile, second and third quartile and top quartile. The size of this data set proved problematic to be processed in R. Therefore, no further analysis was conducted.

## 4.2 Classical Classification Approaches

As noted above, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA),and logistic regression were used. LDA and QDA, when applied directly to the data, yielded only predictions of improvement, so the confusion matrices are not shown here. Both assume that observations are drawn from a Gaussian distribution, which is not true in this case, so various transformations were applied to the data. However, these modifications led to no improvements. While they were somewhat accurate simply due to the skewed nature of the data, they were not useful for our purposes.

Logistic regression does not make these assumptions, and led to more usable data. The confusion matrix below outlines the results of the logistic regression analysis using a standard probability threshold of 50%.

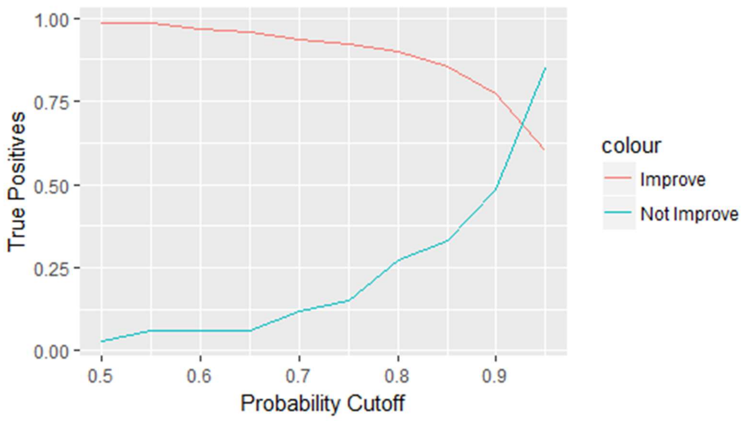
Logistic Regression with Standard Assumptions

| | Not Improve | Improve | Total |
|---|---|---|---|
| Not Improve | 1 | 3 | 4 |
| Improve | 37 | 465 | 502 |
| Total | 38 | 468 | 506 |
| Specificity | | 3% | |
| False-Negative Rate | | 75% | |

This approach showed a fairly high accuracy rating simply due to the volume of participants whose poverty levels improved, but a low specificity. Additional exploration was conducted into the effects of adjustment to the predicted probability in order to identify a threshold that would improve the accuracy for the variable of interest. The plot below provides a

[2] http://www.pep-net.org/sites/pep-net.org/files/typo3doc/pdf/CBMS_training/composite_ind.pdf

comparison of the true positive rates for both participants who improved and those who did not improve when adjusting the probability thresholds used for the logistic regression.



The model was rerun with the optimal threshold determined to be 91%. A confusion matrix of this model is provided below.

Logistic Regression with 91% Probability Threshold

| | Not Improve | Improve | Total |
|---|---|---|---|
| Not Improve | 18 | 115 | 133 |
| Improve | 20 | 353 | 373 |
| Total | 38 | 468 | 506 |
| Specificity | | 47% | |
| False-Negative Rate | | 86% | |

## 4.2 Support Vector Machines

Because the response we are interested in is binary, support vector machines can easily be used as another mechanism for classification. Due to the high dimensionality of the data set when including all indicators, it is difficult to determine what kernel may be most successful without further investigation. Therefore, we will examine the results obtained using the linear kernel, polynomial kernel, and radial kernel. Cross-validation was used to determine the optimal cost parameters, as well as the optimal degree for the polynomial kernel and value of gamma for the radial kernel.

For the linear kernel, the optimal identified value for the cost parameter was 32. The confusion matrix for testing data obtained using the linear SVM is shown below.

| | Not Improve | Improve | Total |
|---|---|---|---|
| Not Improve | 64 | 41 | 105 |
| Improve | 78 | 323 | 401 |
| Total | 142 | 364 | 506 |
| Specificity | | 45% | |
| False-Negative Rate | | 39% | |

For the polynomial kernel, the optimal identified value for the cost parameter was 32, with an optimal degree of 6. The confusion matrix for testing data obtained using the polynomial SVM is shown below.

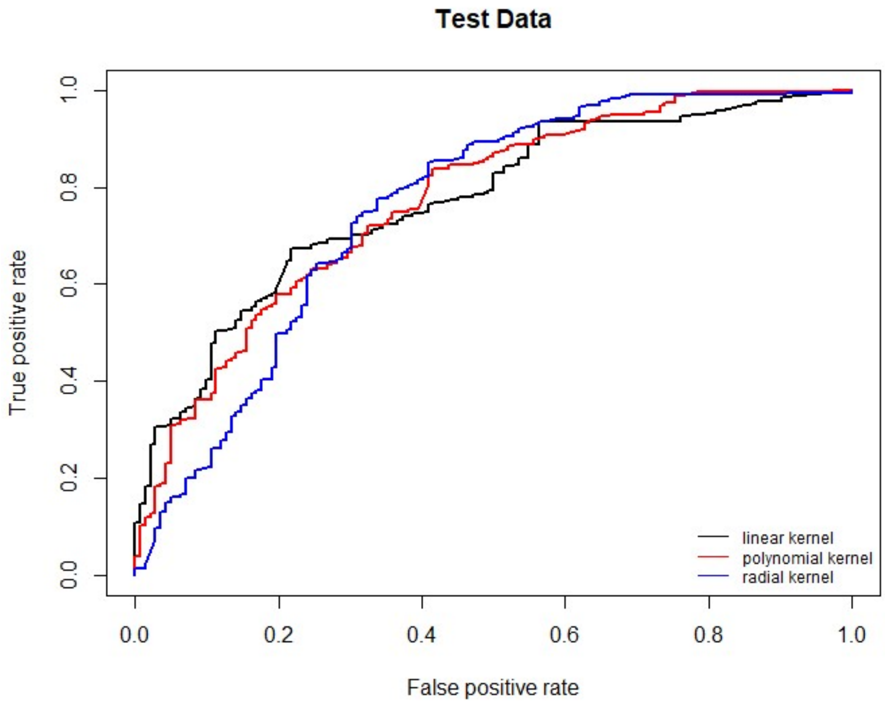
| | Not Improve | Improve | Total |
|---|---|---|---|
| Not Improve | 41 | 13 | 54 |
| Improve | 101 | 351 | 452 |
| Total | 142 | 364 | 506 |
| Specificity | | 29% | |
| False-Negative Rate | | 24% | |

For the radial kernel, the optimal identified value for the cost parameter was 2, with an optimal gamma of 1. The confusion matrix for testing data obtained using the radial kernel is shown below.

| | Not Improve | Improve | Total |
|---|---|---|---|
| Not Improve | 48 | 7 | 55 |
| Improve | 94 | 357 | 451 |
| Total | 142 | 364 | 506 |
| Specificity | | 34% | |
| False-Negative Rate | | 13% | |

Although SVM output is traditionally seen as class labels, these labels are generated from the sign of the numerical scores obtained from the underlying kernels. These scores, also known as the fitted values for each observation, can be used to generate ROC curves. The curve for the most successful model obtained using each kernel and applied to testing data is shown in Figure X.

**Commented [3]:** X



**4.3 Decision Trees**

As with LDA and QDA, using decision trees for classification yielded only predictions of improvement, so the resulting confusion matrix is not included. Aggregate dimension scores were used because trees do not function well with qualitative data, and the methods tried were as follows: pruning to five branches after cross-validating the tree size; bagging; random forests using m = 3 after cross-validation; and boosting using the Brieman (0.5*ln(1-error/error)), Freund (ln(1-error/error)), and Zhu (ln(1-error/error)+ln(#classes - 1)) shrinking parameters.

A single decision tree that was representative of our findings is shown below to highlight the identical responses at each of the leaves.

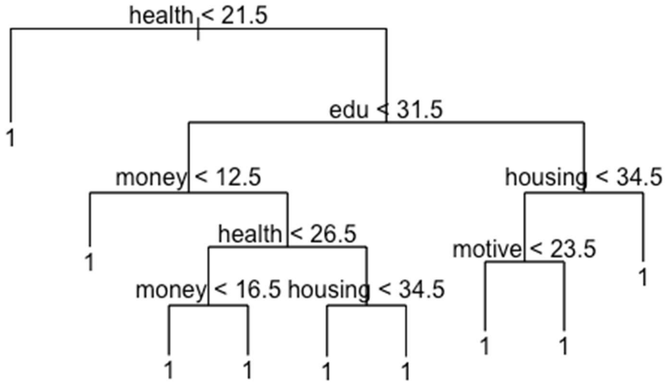


Figure 6. Single decision tree with cutpoints

### 4.4 k-Nearest Neighbors

kNN with k = 1 was able to decently classify the families that improved versus the ones who did not. Accuracy decreased with larger values of k, so those confusion matrices are not included. We evaluated kNN for the data when considering the 50 separate indicators, as well as the neighbors for observations when considering the 6 poverty dimensions.

KNN-50 Indicators

| | Not Improve | Improve | Total |
|---|---|---|---|
| Not Improve | 18 | 14 | 32 |
| Improve | 29 | 445 | 474 |
| Total | 47 | 459 | 506 |
| Specificity | | 38% | |
| False-Negative Rate | | 44% | |

KNN-6 Dimensions

| | Not Improve | Improve | Total |
|---|---|---|---|
| Not Improve | 16 | 3 | 19 |
| Improve | 29 | 458 | 487 |
| Total | 45 | 461 | 506 |
| Specificity | | 36% | |
| False-Negative Rate | | 16% | |

# 5 Discussion

As can be seen in the summary table for models provided below, the overall accuracy of most of the models was fairly good, primarily due to the high proportion of participants who saw improvement in their state of poverty. However, the primary objective of this analysis was to maximize the specificity while also minimizing the false negative rate, suggesting that SVM with a radial kernel and kNN considering the 6 poverty dimensions output the most useful models.

| Rate | Logistic Regression<br>Threshold of 50% | Logistic Regression<br>Threshold of 91% | SVM, Linear | SVM, Polynomial | SVM, Radial | kNN-50 Indicators | kNN-6Dimensions |
|---|---|---|---|---|---|---|---|
| Specificity | 3% | 47% | 45% | 29% | 34% | 38% | 36% |
| False-Negative Rate | 75% | 86% | 39% | 24% | 13% | 44% | 16% |
| Accuracy | 92% | 73% | 76% | 77% | 80% | 92% | 94% |

<u>LDA/QDA/Trees</u>

As mentioned in their respective portions of Section 4, LDA, QDA and classification trees output only predictions of improvement and their confusion matrices were not considered. Each of these three is characterized by making predictions based on the most commonly occurring class given the context of the data point. For LDA and QDA, this is achieved by mimicking the Bayes' classifier, which considers the most likely class for a given observation given its predictor values. For classification trees, this is achieved by considering the most commonly occurring class of observations in that observation's region. Because the true data in our considered test sets was overwhelmingly comprised of observations that saw improvement, it is unsurprising that improvement was seen as the likeliest outcome in either case; nonetheless, it means that these should be not be considered as viable options for our purposes.

<u>Logistic Regression</u>

Logistic regression was accurate only in cases that yielded low specificity; when the threshold was changed to improve specificity, the accuracy was significantly impacted and the false-negative rate rose. Therefore, this approach is not well-suited to our purposes. Logistic regression is not recommended in contexts where the classes are well-separated; here, that would be mean that it performs poorly when individuals who show no improvement from participation in the Stoplight program are significantly distinct from those who do show improvement. The high-dimensionality of the data set makes it difficult to visually show that the classes are well-separated, but the success of support vector machines, which are similar to logistic regression but perform better with well-separated classes, supports this notion.

## Support Vector Machines

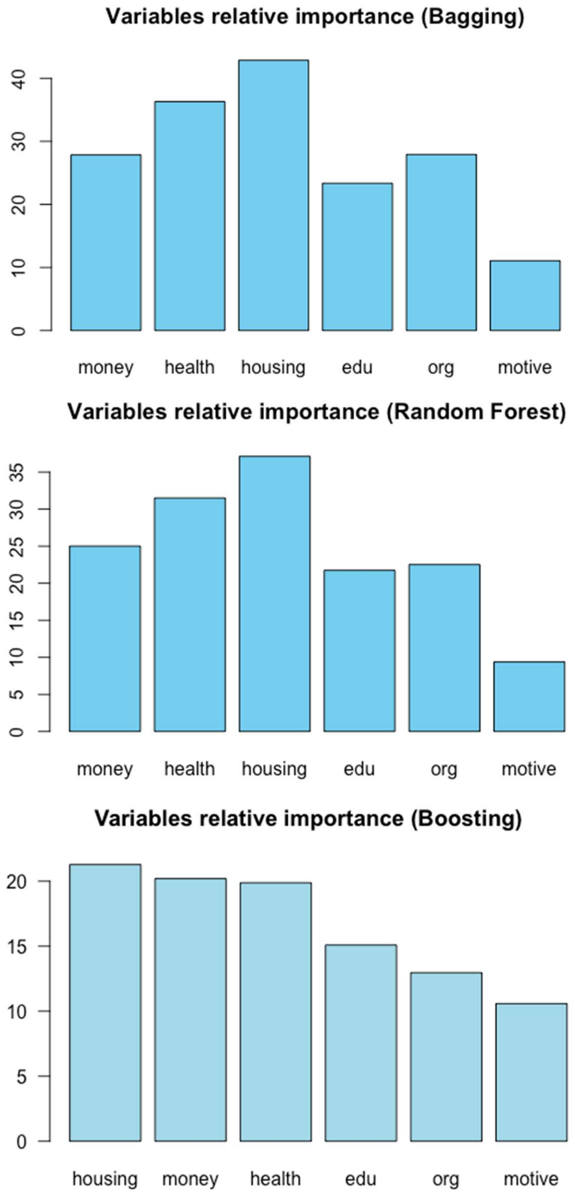
The accuracies for each kernel are extremely similar, which is, in general, unexpected due to the high differences in approach between the three kernels. Because the data is classified into only three levels for each dimension, the data is not highly separated. Subsequently, due to the overlapping nature of the data, each kernel saw more than half of the data be used as support vectors; for the radial kernel, nearly 90% of the data served as support vectors. This serves to reduce variance, but can lead to potentially high bias, and indicates that our evaluation would be improved if the data were more separated. However, as discussed above, the data is well-separated enough to make application of logistic regression difficult. We can conclude, then, that SVM might might work better the data were more well-separated, or that logistic regression might see improved results if the data were less well-separated, but the data currently exists in a context where neither functions optimally.

## KNN

The k-nearest neighbors approach with low values of k is predicated on similarity between the characteristics of individuals. In our context, with k = 1, it means that we are predicting the outcome of participation in the Stoplight program solely based on the outcome for the individual with the most similar characteristics to each given test observation.

## Trees

Tree models were not useful for classification, but were capable of providing information for use in inference by considering which variables played the most important role in reducing variation. Using random forest, bagging and boosting approaches, the relative importance of the variables was fairly consistent. Housing was rated as having the highest level of important followed by health and money. The plot for random forests had m = 3, and the plot for boosting had the Brieman learning coefficient.



Figures 7-9. Importance plots for bagging, random forests (m = 3), and boosting (Brieman coefficient), using quantitative data aggregated by category.

# 6 Conclusion

Overall, the support vector machine with a radial kernel and the kNN model using the 6 poverty dimensions best balanced the objectives of the model: Maximizing specificity and minimizing the false negative rate. However, kNN with k = 1 is far easier to interpret, as it is relatively easy to identify the past participants with the most similar dimension scores, and is therefore the method that we recommend. The simplicity inherent in kNN with k = 1 is also useful, as it is a much easier concept to explain to program personnel who may have less experience with statistical applications.

The importance plots suggest that perhaps allocating additional resources/placing additional emphasis on housing and health related issues may help to improve the likelihood that a participant is successful at improving their overall poverty score.

References

1 Fundación Paraguaya. 2015. *Growing with "El Mejor": Impact Measurement Report*.

2 Husson, Francois, Sebastien Le, and Jérôme Pagès. 2017. *Exploratory Multivariate Analysis by Example Using R*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. http://factominer.free.fr/bookV2/index.html.